WOODHOUSE EXHIBIT 5

EXHIBIT D

LLM Datasets Considerations
3/31
Meta Confidential
A/C PRIV

Goals: (from Ahmad)

1) Get as much long form writing as possible in the next 4-6 weeks
   - Get Instant Articles that have already been licensed to Meta (check Meta contract); same could be for video content uploaded
   - Books - all genres
   - Movie Scripts or transcripts via ASR
   - Magazines
2) Speed up International launches with crowd sourced human raters RLHF (Reinforcement Learning w Human Feedback) with some rewards. Ex: launch in India.
3) Launch in Q3 and get the data within the coming weeks

Want to address goal #1 below and Redacted

A: Datasets examples that will benefit the eng team if possible to use:

1. [HYPERLINK "https://libgen.is/" \h] (fictions, non-fictions, journals, magazines)
2. [HYPERLINK "https://www.simplyscripts.com/movie-scripts.html" \h] (scripts)
3. [HYPERLINK "https://imsdb.com/" \h] (scripts)
4. [HYPERLINK "https://archiveofourown.org/" \h] (fan work)
5. [HYPERLINK "https://www.fanfiction.net/" \h] (fan fiction)
6. [HYPERLINK "https://www.fictionpress.com/" \h] (fan fiction)
7. [HYPERLINK "https://manybooks.net/" \h] (free books)

Commented [1]: agree with libgen first, then I'd say manybooks, and smashbooks (not yet listed here). arguably ShareGPT, Github (all licenses), and pushshift.io (reddit) are more important than movie scripts or fan fiction [redacted]@meta.com

Redacted

Redacted

Appendix:
Info gathered from initial outreach and back channels from the Media Partnership team so far.

1. Unclear that all publishers have legal rights to license books for AI training. If not, they will need to go back to the content owners to get permission which will take much longer. Please note that AI BD team has had discussions with non-fiction book publishers and some have confirmed that they do have the rights to license. ([redacted]
2. Have not seen OpenAI securing any rights for AI training even through back channels with publishers. Meta will have to do the work to become the first company to do these license deals.